D242–D251 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 21 November 2011*
doi:10.1093/nar/gkr1064

ELM—the database of eukaryotic linear motifs

Holger Dinkel[1], Sushama Michael[1], Robert J. Weatheritt[1], Norman E. Davey[1], Kim Van Roey[1], Brigitte Altenberg[1], Grischa Toedt[1], Bora Uyar[1], Markus Seiler[1], Aidan Budd[1], Lisa Jödicke[1], Marcel A. Dammert[1], Christian Schroeter[1], Maria Hammer[1], Tobias Schmidt[1], Peter Jehl[1], Caroline McGuigan[1], Magdalena Dymecka[2], Claudia Chica[3], Katja Luck[4], Allegra Via[5], Andrew Chatr-aryamontri[6], Niall Haslam[7], Gleb Grebnev[7], Richard J. Edwards[8], Michel O. Steinmetz[9], Heike Meiselbach[10], Francesca Diella[1,11] and Toby J. Gibson[1,*]

[1]Structural and Computational Biology, European Molecular Biology Laboratory, Heidelberg, Germany, [2]Laboratory of Bioinformatics and Systems Biology, M. Sklodowska-Curie Cancer Center and Institute of Oncology, WK Roentgena 5, 02-781 Warsaw, Poland, [3]Genoscope (CEA – Institut de Génomique), 2 rue Gaston Cremieux CP5706, 91057 Evry, [4]Group Oncoproteins, Unité CNRS-UDS UMR 7242, Institut de Recherche de l'Ecole de Biotechnologie de Strasbourg, 1, Bd Sébastien Brant, BP 10413, 67412 Illkirch – Cedex, France, [5]Biocomputing Group, Department of Physics, Sapienza University of Rome, P.le Aldo Moro 5, Rome, Italy, [6]School of Biological Sciences, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JR, UK, [7]School of Medicine and Medical Science, University College, Dublin, Ireland, [8]Centre for Biological Sciences, Institute for Life Sciences, University of Southampton, UK, [9]Biomolecular Research, Paul Scherrer Institut, CH-5232 Villigen PSI, Switzerland, [10]Bioinformatik, Institut für Biochemie, Friedrich-Alexander-Universität, Fahrstraße 17, 91054 Erlangen-Nürnberg and [11]Molecular Health GmbH Belfortstr. 2, 69115 Heidelberg, Germany

Received September 13, 2011; Revised and Accepted October 27, 2011

ABSTRACT

Linear motifs are short, evolutionarily plastic components of regulatory proteins and provide low-affinity interaction interfaces. These compact modules play central roles in mediating every aspect of the regulatory functionality of the cell. They are particularly prominent in mediating cell signaling, controlling protein turnover and directing protein localization. Given their importance, our understanding of motifs is surprisingly limited, largely as a result of the difficulty of discovery, both experimentally and computationally. The Eukaryotic Linear Motif (ELM) resource at http://elm.eu.org provides the biological community with a comprehensive database of known experimentally validated motifs, and an exploratory tool to discover putative linear motifs in user-submitted protein sequences. The current update of the ELM database comprises 1800 annotated motif instances representing 170 distinct functional classes, including approximately 500 novel instances and 24 novel classes. Several older motif class entries have been also revisited, improving annotation and adding novel instances. Furthermore, addition of full-text search capabilities, an enhanced interface and simplified batch download has improved the overall accessibility of the ELM data. The motif discovery portion of the ELM resource has added conservation, and structural attributes have been incorporated to aid users to discriminate biologically relevant motifs from stochastically occurring non-functional instances.

INTRODUCTION

Short linear motifs (SLiMs, LMs or MiniMotifs) are regulatory protein modules characterized by their compact interaction interfaces (the affinity and specificity determining residues are usually encoded between 3 and 11 contiguous amino acids (1)) and their enrichment in natively unstructured, or disordered, regions of proteins (2). As a result of limited intermolecular contacts with their interaction partners, SLiMs bind with relatively

*To whom correspondence should be addressed. Tel: +49 (0) 6221 3878398; Fax: +49 (0) 6221 387517; Email: gibson@embl-heidelberg.de

low affinity (in the low-micromolar range), an advantageous attribute for use as transient, conditional and tunable interactions necessary for many regulatory processes. Due to the limited number of mutations necessary for the genesis of a novel motif, SLiMs are amenable to convergent evolution, functioning as a driver of network evolution by adding novel interaction interfaces, and thereby new functionality, to proteins. This evolutionary plasticity facilitates the rapid proliferation within a proteome, and as a result, motif use is ubiquitous in higher eukaryotes.

SLiMs play an important role for many regulatory processes such as signal transduction, protein trafficking and post-translational modification (3,4). Their importance to the correct functionality of the cell is also reflected by the outcome of motif deregulation. For example, point mutations in SLiMs have been shown to lead severe pathologies such as 'Noonan-like syndrome' (5), 'Liddle's syndrome' (6) or 'Retinitis pigmentosa' (7). Furthermore, mimicry of linear motifs by viruses to hijack their hosts' existing cellular machinery plays an important role in many viral life cycles (8). However, despite their obvious importance to eukaryotic cell regulation, our understanding of SLiM biology is relatively limited, and it has been suggested that, to date, we have only discovered a small portion of the human motifs (9).

Several resources are devoted to the annotation and/or detection of SLiMs [Prosite (10), MiniMotifMiner (11) and Scansite (12)]. Here, we report on the 2012 status of the Eukaryotic Linear Motif database.

THE ELM RESOURCE

The ELM initiative (http://elm.eu.org) has focused on gathering, storing and providing information about short linear motifs since 2003. It was established as the first manually annotated collection of SLiM classes and as a tool for discovering linear motif instances in proteins (13). As it was mainly focused on the eukaryotic sequences, it was termed the Eukaryotic Linear Motif resource, usually shortened to ELM. The ELM resource consists of two applications: the ELM database of curated motif classes and instances, and the motif detection pipeline to detect putative SLiM instances in query sequences. In the ELM database, SLiMs are annotated as 'ELM classes', divided into four 'types': cleavage sites (CLV), ligand binding sites (LIG), sites of post-translational modification (MOD) and subcellular targeting sites (TRG) (Table 1). Currently, the ELM database contains 170 linear motif classes with more than 1800 motif instances linked to more than 1500 literature references (Table 1). Each class is described by a regular expression capturing the key specificity and affinity determining amino acid residues. A regular expression is a computer-readable term for sequence annotation and is used by the ELM motif detection pipeline to scan proteins for putative instances of annotated ELM classes. The search form for sequence input is shown in Figure 1, while the results page showing the putative and annotated instances is illustrated in Figure 2.

The ELM resource is powered by a PostgreSQL relational database for data storage and a PYTHON web framework for data retrieval/visualization. The main tables within the database contain information about ELM classes, ELM instances, sequences, references, taxonomy and links to other databases [the database structure is described in greater detail in (14)].

New ELM classes

Since the last release (14), 24 new ELM classes have been added to the ELM database (Table 1) and several more have been updated. One of the newly annotated motif classes is the AGC kinase docking motif (LIG_AGCK_PIF), consisting of three distinct classes. It is present in the non-catalytic C-terminal tail of AGC kinases that constitute a family of serine/threonine kinases consisting of 60 members that regulate critical processes, including cell growth and survival. Deregulation of these enzymes is a causative factor in different diseases such as cancer and diabetes. The motif interacts with the PDK1 Interacting Fragment (PIF) pocket in the kinase domain of AGC kinases. It mediates intramolecular binding to the PIF pocket, serving as a *cis*-activating module together with other regulatory sequences in the C-tail. Interestingly, in some kinases the motif also acts as a PDK1 docking site that *trans*-activates PDK1, which itself lacks the regulatory C-tail, by interacting with the PDK1 PIF pocket. PDK1 in turn will phosphorylate and activate the docked kinase. Other novel classes (Table 2) include phosphodegrons, which are important mediators of phosphorylation-dependent protein destruction, and the LYPxL motif, which is involved in endosomal

Table 1. Summary of data stored in the ELM database[a]

Number of functional site entries		ELM motif classes		ELM motif instances		Links to PDB structures	GO terms		Pubmed links	
Totals	115		170		1840	195		340		1561
By category		LIG	111	Human	1004					
		MOD	30	Mouse	160		Biological process	173	From ELM motif	787
		TRG	21	Rat	102					
		CLV	8	Fly	67		Cell compartment	74	From instance	1071
				Yeast	90					
				Other	417		Molecular function	93		

[a]As of October 2011.

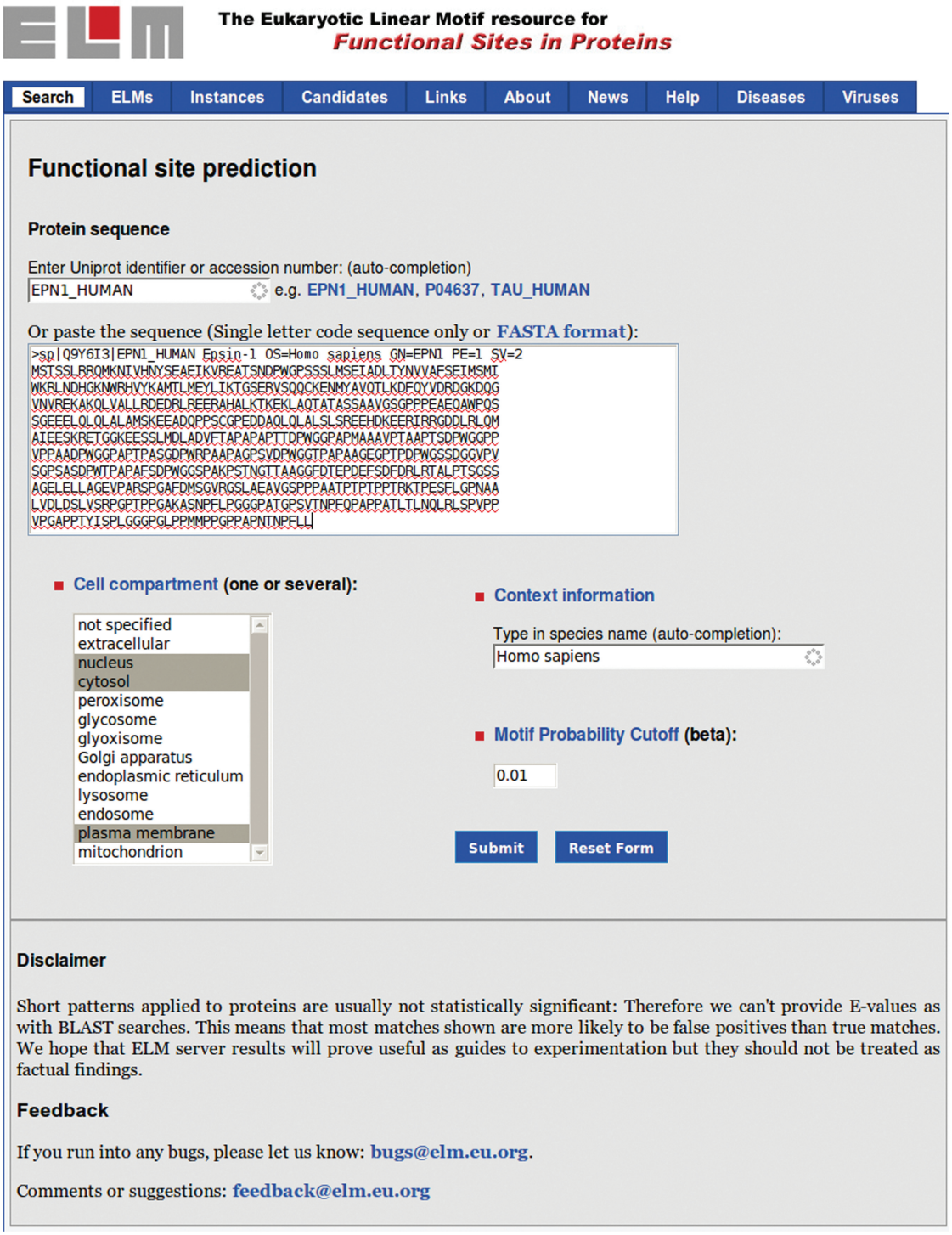


Figure 1. ELM start page. The user can submit a query sequence to the motif detection pipeline either as UniProt accession number or in FASTA format. Filtering criteria such as taxonomic range or cellular compartment should be activated to limit the resulting list of SLiM instances.

sorting of membrane proteins but is also implicated in retrovirus budding.

New ELM instances

Annotated ELM instances serve as representative examples of the respective ELM class. They are also invaluable for the computational analysis and classification of motifs (15). Therefore, special emphasis has been put on the curation of more than 500 novel ELM instances (in 40 different classes) by scanning and annotating more than 400 articles. The number of protein databank (PDB) entries annotated have been increased to 195 (Table 1), meaning that for ~10% of all instances there is a 3D

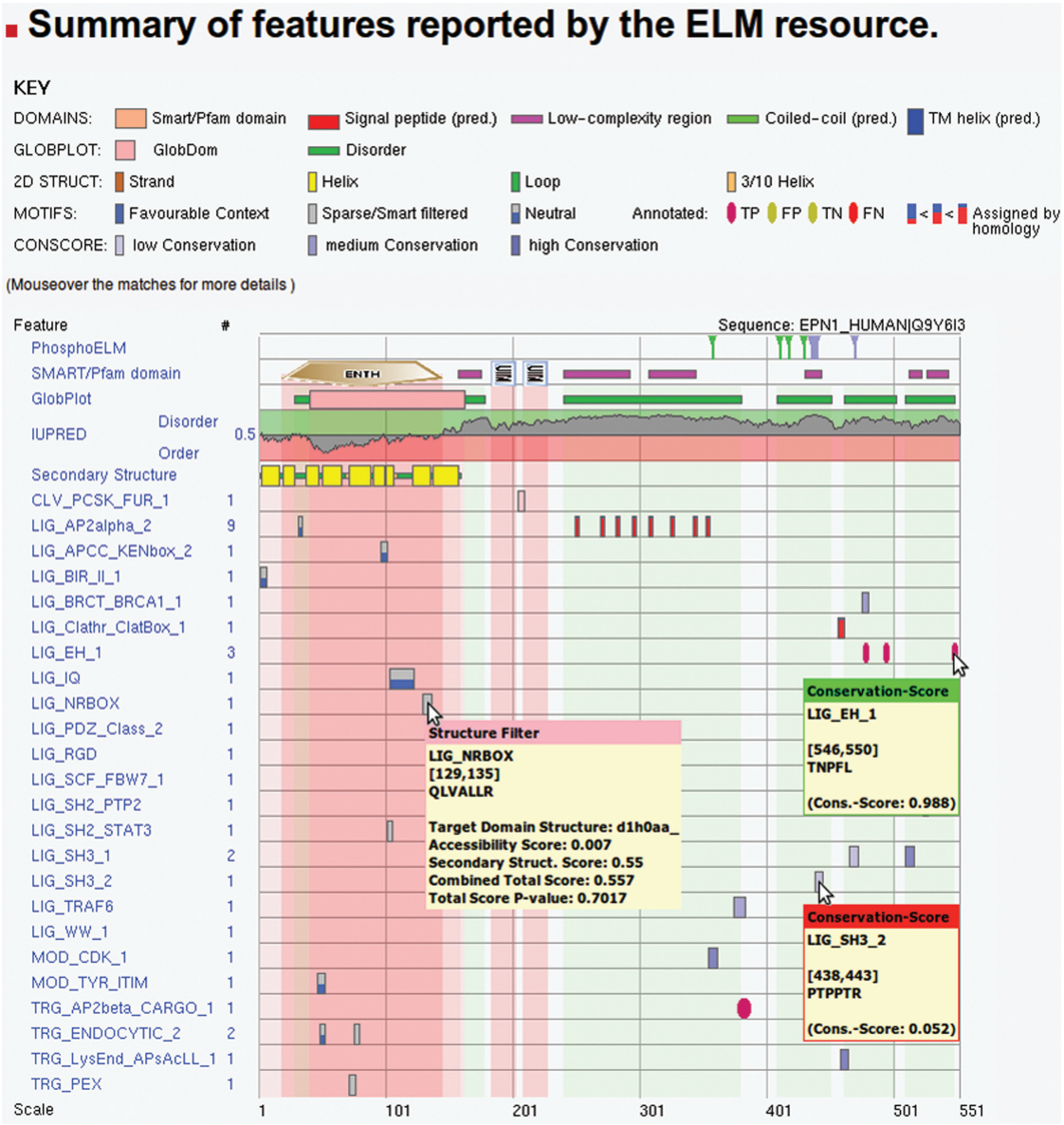


Figure 2. ELM motif detection pipeline output page. The top legend explains the different colors/symbols used. The graphical output of ELM concentrates the output of multiple sequence classification algorithms; phosphorylation sites from Phospho.ELM, protein domains detected by SMART/Pfam, disorder predictions by GlobPlot and IUPred and secondary structure (18). The lower part contains the annotated and putative ELM instances for the given protein sequence (Epsin1, UniProt accession Q9Y6I3). The background is colored according to the structural information available. Each box represents one ELM instance, the color of which indicates the likelihood that this instance is functional: grey instances are buried within structured regions, while shades of blue represent instances outside of structured regions and hint on sequence conservation, with pale blue representing weak sequence conservation and dark blue indicating strong sequence conservation. Red ellipses or boxes mark instances that are annotated in the query sequence or a homologous sequence, respectively.

protein structure annotated, giving more detailed information about the biological context of the respective motif.

NEW FEATURES

The ELM website at http://elm.eu.org can be used in two ways: first, as a front-end to explore the ELM database of curated ELM classes and instances, and second, to run the motif detection pipeline to detect putative SLiM instances in query sequences. Both interfaces have been improved with the most notable changes listed below.

User interface

The database user interface, having been stable for many years, has been overhauled and replaced by a novel interface introducing several new features (Figure 1). Up-to-date web technologies have been used to improve the general user experience: the PYTHON framework DJANGO (http://www.djangoproject.com) dynamically creates and serves all HTML pages, while JavaScript was used to make the whole site more interactive and thus improve the user experience. In particular, the ELM detail pages (Figure 3), which hold the most

Table 2. List of novel ELM classes[a]

Identifier	Description
LIG_Actin_WH2_1 LIG_Actin_WH2_2 LIG_Actin_RPEL_3	Motifs, present in proteins in several repeats, which mediate binding to the hydrophobic cleft created by subdomains 1 and 3 of G-actin
LIG_AGCK_PIF_1 LIG_AGCK_PIF_2 LIG_AGCK_PIF_3	The AGCK docking motif mediates intramolecular interactions to the PDK1 Interacting Fragment (PIF) pocket, serving as a *cis*-activating module
LIG_BIR_II_1 LIG_BIR_III_1 LIG_BIR_III_2 LIG_BIR_III_3 LIG_BIR_III_4	IAP-binding motifs are found in pro-apoptotic proteins and function in the abrogation of caspase inhibition by inhibitor of apoptosis proteins in apoptotic cells
LIG_eIF4E_1 LIG_eIF4E_2	Motif binding to the dorsal surface of eIF4E
LIG_EVH1_3	A proline-rich motif binding to EVH1/WH1 domains of WASP and N-WASP proteins
LIG_HCF-1_HBM_1	The DHxY Host Cell Factor-1 binding motif interacts with the N-terminal kelch propeller domain of the cell cycle regulator HCF-1
LIG_Integrin_isoDGR_1	Present in proteins of extracellular matrix which upon deamidation forms biologically active isoDGR motif which binds to various members of integrin family
LIG_LYPXL_L_2 LIG_LYPXL_S_1	The LYPxL motif binds the V-domain of Alix, a protein involved in endosomal sorting
LIG_PAM2_1	Peptide ligand motif that directly interacts with the MLLE/PABC domain found in poly(A) binding proteins and HYD E3 ubiquitin ligases
LIG_PIKK_1	Motif located in the C terminus of Nbs1 and its homologous interacting with PIKK family members
LIG_Rb_pABgroove_1	The LxxLFD motif binds in a deep groove between pocket A and pocket B of the Retinoblastoma protein
LIG_SCF_FBW7_1 LIG_SCF_FBW7_2	The TPxxS phospho-dependent degron binds the FBW7 F box proteins of the SCF (Skp1-Cullin-Fbox) complex
LIG_SPAK-OSR1_1	SPAK/OSR1 kinase binding motif acts as a docking site which aids the interaction with their binding partners including the upstream activators and the phosphorylated substrates

[a]As of October 2011.

important information about each ELM class including references, regular expression, taxonomic distribution and gene ontology terms (Table 3), have been updated by annotating the protein domain interacting with the respective motif. Where available, a 3D model of representative protein databank structures of linear motif interactions was added to the ELM detail page (Figure 3, top right).

To cope with the increasing amount of annotated classes as well as instances, a novel query interface was introduced to assist the user in finding information of interest. The ELM browser (Figure 4) now features a search interface for free text search. In addition, the search results can also be filtered and reordered using buttons (Figure 4, left side) and table headers, respectively, and be downloaded as tab-separated values (TSV).

Further, improvements to the ELM database include revising the experimental methods used for annotation by using a standardized methods vocabulary [in sync with PSI-MI ontology (16,17)].

A candidate page has been introduced to display novel ELM classes that have not yet been annotated in detail or are currently undergoing annotation. We invite researchers to send us their feedback and expert opinion on these classes and to contribute novel motif classes that will be added to the candidate page and ultimately be turned into full ELM classes (Figure 5). Minimum requirements are at least one literature reference as well as a short description. In addition, a draft regular expression or a 3D structure showing the relevant interaction would also be helpful. Currently, the number of possible ELM classes on this candidate list (awaiting further annotation) exceeds the number of completely annotated classes, indicating the great demand for further annotation.

Graphical representation of sequence search

The ELM motif detection pipeline scans protein sequences for matches to the regular expressions of annotated ELM classes (Figure 2). The query output combines these putative instances with information from the database (annotated ELM instances) as well as predictions from different algorithms/filters. The ELM resource employs a structural filter (18) to highlight and mask secondary structure elements, as well as SMART (19) to detect protein domains. Furthermore, an additional disorder prediction algorithm (IUPred) (20) has been included to predict ordered/disordered regions within the protein. IUPred uses a cutoff of 0.5 to classify a sequence region as either structured or disordered, with values above this threshold corresponding to disorder, highlighted in green background and lower values indicating structured regions, displayed in red background in the output graph. Disorder and domain information is combined by

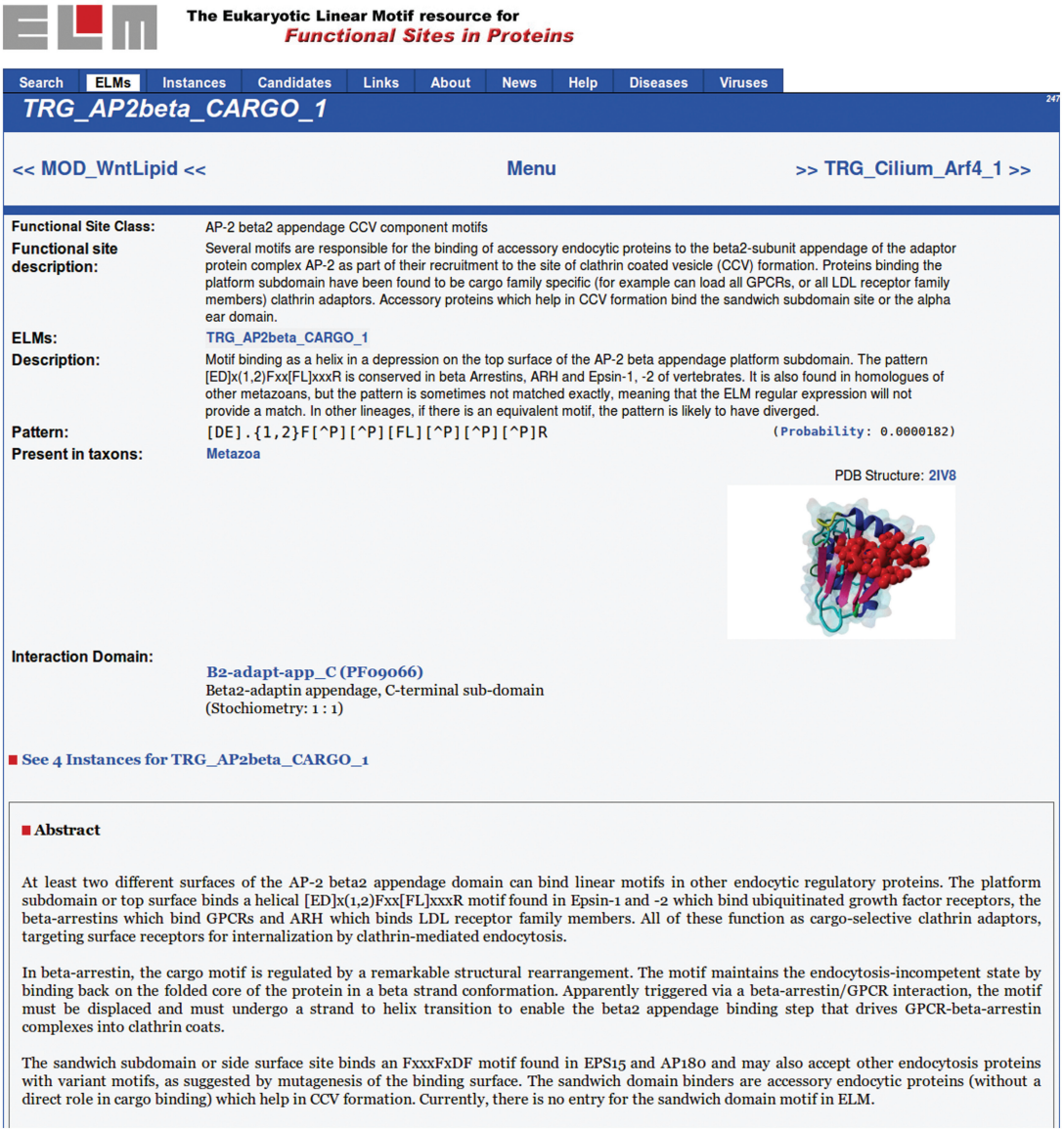


Figure 3. ELM detail page showing information about the ELM class TRG_AP2beta_CARGO_1.

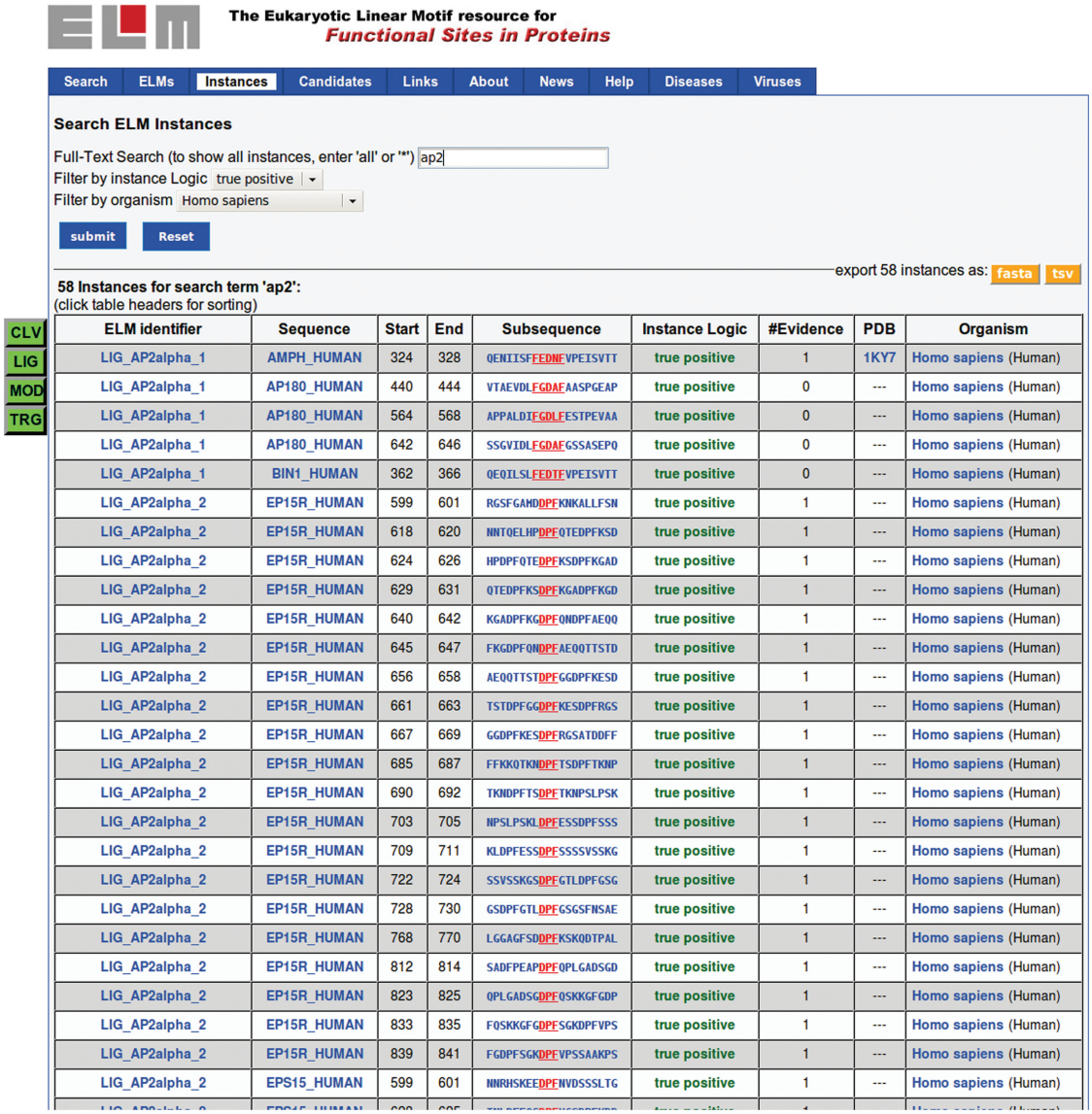
Table 3. Main cellular compartments used in ELM annotation

Count	GO Id	GO term
98	GO:0005829	Cytosol
69	GO:0005634	Nucleus
17	GO:0005576	Extracellular
12	GO:0005794	Golgi apparatus
10	GO:0005886	Plasma membrane
9	GO:0009898	Internal side of plasma membrane
9	GO:0005783	Endoplasmic reticulum
6	GO:0005739	Mitochondrion
5	GO:0005643	Nuclear pore
5	GO:0045334	Clathrin-coated endocytic vesicle

background coloring to highlight structured regions within the protein, which allows inspection of SLiMs that reside at domain boundaries and emphasizes motifs in disordered regions.

The conservation of linear motifs can help in assessing the functional relevance of putative instances, with functional instances showing higher overall sequence conservation than non-functional ones (21). Therefore, sequence conservation of the query protein is calculated using a tree-based conservation scoring method (22) and highlighted in the graphical output. Here, lighter shades of blue represent low conservation while dark blue shading corresponds to high-sequence conservation. The actual conservation score can be inspected by moving the mouse over the respective ELM instance (Figure 2).

The functionality of linear motifs can be modulated by modifications such as phosphorylation (23,24). To enable the user to investigate phosphorylation data in the context of putative linear motif instances, phosphorylation annotations from the Phospho.ELM resource (25) have been added to the graphical output (Figure 2, top row).



ELM identifier	Sequence	Start	End	Subsequence	Instance Logic	#Evidence	PDB	Organism
LIG_AP2alpha_1	AMPH_HUMAN	324	328	QENIISFFEDNEVPEISVTT	true positive	1	1KY7	Homo sapiens (Human)
LIG_AP2alpha_1	AP180_HUMAN	440	444	VTAEVDLFGDAFAASPGEAP	true positive	0	---	Homo sapiens (Human)
LIG_AP2alpha_1	AP180_HUMAN	564	568	APPALDIFGDLFESTPEVAA	true positive	0	---	Homo sapiens (Human)
LIG_AP2alpha_1	AP180_HUMAN	642	646	SSGVIDLFGDAFGSSASEPQ	true positive	0	---	Homo sapiens (Human)
LIG_AP2alpha_1	BIN1_HUMAN	362	366	QEQILSLFEDTFVPEISVTT	true positive	0	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	599	601	RGSFGAMDDPFKNKALLFSN	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	618	620	NNTQELHPDPFQTEDPFKSD	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	624	626	HPDPFQTEDPFKSDPFKGAD	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	629	631	QTEDPFKSDPFKGADPFKGD	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	640	642	KGADPFKGDPFQNDPFAEQQ	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	645	647	FKGDPFQNDPFAEQQTTSTD	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	656	658	AEQQTTSTDPFGGDPFKESD	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	661	663	TSTDPFGGDPFKESDPFRGS	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	667	669	GGDPFKESDPFRGSATDDFF	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	685	687	FFKKQTKNDPFTSDPFTKNP	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	690	692	TKNDPFTSDPFTKNPSLPSK	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	703	705	NPSLPSKLDPFESSDPFSSS	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	709	711	KLDPFESSDPFSSSSVSSKG	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	722	724	SSVSSKGSDPFGTLDPFGSG	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	728	730	GSDPFGTLDPFGSGSFNSAE	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	768	770	LGGAGFSDDPFKSKQDTPAL	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	812	814	SADFPEAPDPFQPLGADSGD	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	823	825	QPLGADSGDPFQSKKGFGDP	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	833	835	FQSKKGFGDPFSGKDPFVPS	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EP15R_HUMAN	839	841	FGDPFSGKDPFVPSSAAKPS	true positive	1	---	Homo sapiens (Human)
LIG_AP2alpha_2	EPS15_HUMAN	599	601	NNRHSKEEDPFNVDSSSLTG	true positive	1	---	Homo sapiens (Human)

Figure 4. ELM instances browse page. A full-text search (here, search term used was 'AP2', filtering for 'true positive' instances in taxon '*Homo sapiens*', yielding 58 instances) assists in finding annotated instances. A search can be restricted to a particular taxonomy or instance logic (top) or ELM class type (buttons on the left). The list can also be exported to TSV or FASTA format for further processing.

The phosphorylated residues are highlighted in different colors (serine: green, threonine: blue, tyrosine: red); each phosphorylation site is linked to a page showing detailed information about the respective modification site from the manually curated data set of the Phospho.ELM resource.

VIRAL INSTANCES

The importance of the short linear motifs in virus–host interactions makes the ELM resource an important tool for the viral research community. For example, Cruz *et al.* (26) analyzed a protein phosphatase 1 (PP1) docking motif in 'protein 7' of transmissible gastroenteritis virus using the ELM class LIG_PP1. This conserved sequence motif mediates binding to the PP1 catalytic subunit, a key regulator of the cellular antiviral defense mechanisms, and is also found in other viral proteomes, suggesting that it might be a recurring strategy to counteract the hosts' defense against RNA viruses by dephosphorylating eukaryotic translation initiation factor 2α and ultimately ribonuclease L.

To reflect our increasing awareness of viral motifs (8), special focus has been attributed to the annotation of viral instances in the ELM database: in the latest release, more than 200 novel ELM instances found in 84 different viral taxons have been added. The notion of viruses abusing existing SLiMs in their hosts is demonstrated by viral instances being annotated alongside instances in their hosts' proteins. For example, the ELM class LIG_PDZ_Class_1 contains 12 instances in human proteins but has recently been expanded with 5 instances from 5 different human pathogenic virus proteins.

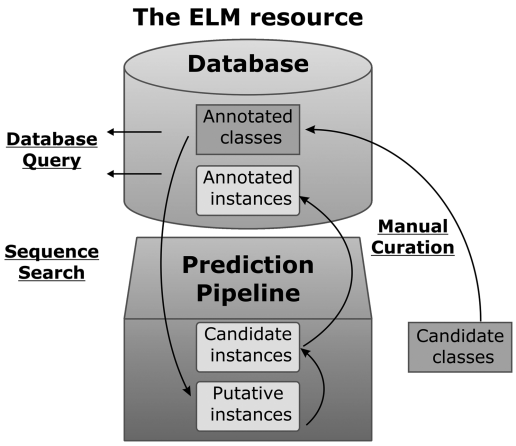


Figure 5. Schema of the ELM resource and data life cycle. Annotated ELM classes, and instances thereof, can be searched by database query. Via sequence search by the motif detection pipeline, annotated ELM classes yield putative instances in query sequences. By adding experimental evidence and references, these putative instances become candidate instances for annotation, and, with further curation, ultimately become fully annotated instances.

LINEAR MOTIFS AND DISEASES

The importance of SLiMs is further corroborated by the occurrence of pathologies that are caused by mutations that either mutate existing linear motifs or create novel linear motifs (of undesired function) (27). Examples include 'Usher's syndrome' (28), 'Liddle's Syndrome' (6) or 'Golabi-Ito-Hall Syndrome' (29). The developmental disorder 'Noonan Syndrome' can be caused by mutations in Raf-1 that abrogate the interaction with 14-3-3 proteins mediated by corresponding SLiMs and thereby deregulate the Raf-1 kinase activity (30) (the Raf-1 protein sequence features two LIG_14-3-3_1 binding sites that are annotated at 256-261 and 618-623 in the ELM resource). A related disease, 'Noonan-like Syndrome', is caused by an S to G mutation at position 2 of the SHOC2 protein, creating a novel myristoylation site (annotated as ELM class MOD_NMyristoyl). This irreversible modification results in aberrant targeting of SHOC2 to the plasma membrane and impaired translocation to the nucleus upon growth factor stimulation (5). More information about the implication of short linear motifs on diseases is collected at http://elm.eu.org/infos/diseases.html.

APPLICATION OF THE ELM RESOURCE

By providing a high-quality, manually curated data set of linear motif classes with experimentally validated SLiM instances, the ELM database has proven to be invaluable to the community: small-scale (single protein) analyzes benefit from the detailed annotation of each ELM class in attributing novel features to proteins of interest. By using *in vitro* and *in vivo* studies, von Nandelstadh *et al.* (31) could validate a PDZ class III motif, detected by ELM at the carboxy terminus of myotilin and the FATZ (calsarcin/myozenin) families. This evolutionarily conserved carboxy-terminal motif mediates binding to PDZ domains of ZASP/Cypher and other Enigma family members (ALP, CLP-36 and RIL) and disruption of these interactions results in myofibrillar myopathies (32). Additionally, ELM annotations can contribute to high-throughput screenings (33) as well as development of novel algorithms (34–36), methods (37) and databases (38). Furthermore, the highly curated data of the ELM resource are used as a benchmarking data set to evaluate the accuracy of prediction algorithms (21,39,40).

For any such analysis, the user should be aware that many matches to ELM regular expressions are false positives. Before conducting experiments based on ELM results, it is strongly advisable to check if a motif match is conserved, exposed in a cell compartment in which the motif is known to be functional. The ELM resource applies several filters to provide the user with such information that should ideally also be supported by the experimental evidence.

SUMMARY

The importance of SLiMs is highlighted by the growing number of instances with relevance to diseases or viruses. Yet, despite their importance and abundance, our understanding of linear motifs is still limited. This is mainly owing to the fact that they are still quite difficult to predict computationally and to investigate experimentally (3,41,42). By better understanding the biology of linear motifs, we hope to increase our insight into diseases and viruses (and vice versa). The ELM resource tries to aid the researcher in the search for putative SLiM instances by providing a feature-rich toolset for sequence analysis. Consequently, with the aforementioned additions and changes, we hope that the ELM resource continues to be a valuable asset to the community.

ACKNOWLEDGEMENTS

The authors would like to thank the users of the ELM resource as well as all colleagues, contributors and annotators of the ELM resource.

FUNDING

EMBL international PhD program (to R.J.W.); EMBL Interdisciplinary PostDoc fellowship (EIPOD to N.E.D.); NGFN framework by the Federal Government Department of Education and Science [FKZ01GS0862 (DiGtoP) to M.S. and M.H.]; European Community's Seventh Framework Programme FP7/2009 (SysCilia) (241955 to G.T.) and (SyBoSS) (242129 to K.V.R.); Polish Ministry of Science and Higher Education within Iuventus Plus project (IP2010-0483-70 to M.D.); Biotechnology and Biological Sciences Research Council (BB/F010486/1 to A.C.); Région Alsace and Collège Doctoral Européen (to K.L.); Science Foundation Ireland (08/IN.1/B1864 to G.G.); BBSRC New Investigator Award (BB/I006230/1 to R.J.E.); German Research Foundation (SFB796 Project A2 to H.M.); grants from the Swiss National Science Foundation (to M.O.S.). Funding for open access charge: EMBL.

Conflict of interest statement. None declared.

REFERENCES

1. Davey,N.E., Van Roey,K., Weatheritt,R.J., Toedt,G., Uyar,B., Altenberg,B., Budd,A., Diella,F., Dinkel,H. and Gibson,T.J. (2011) Attributes of short linear motifs. *Mol Biosyst*., September 12 (doi:10.1039/c1mb05231d; epub ahead of print).
2. Fuxreiter,M., Tompa,P. and Simon,I. (2007) Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, **8**, 950–956.
3. Diella,F., Haslam,N., Chica,C., Budd,A., Michael,S., Brown,N.P., Trave,G. and Gibson,T.J. (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front. Biosci*., 6580–6603.
4. Gibson,T.J. (2009) Cell regulation: determined to signal discrete cooperation. *Trends Biochem. Sci*., **34**, 471–482.
5. Cordeddu,V., Di Schiavi,E., Pennacchio,L.A., Ma'ayan,A., Sarkozy,A., Fodale,V., Cecchetti,S., Cardinale,A., Martin,J., Schackwitz,W. *et al*. (2009) Mutation of SHOC2 promotes aberrant protein N-myristoylation and causes Noonan-like syndrome with loose anagen hair. *Nat. Genet*., **9**, 1022–1026.
6. Furuhashi,M., Kitamura,K., Adachi,M., Miyoshi,T., Wakida,N., Ura,N., Shikano,Y., Shinshi,Y., Sakamoto,K., Hayashi,M. *et al*. (2005) Liddle's syndrome caused by a novel mutation in the proline-rich PY motif of the epithelial sodium channel beta-subunit. *J. Clin. Endocrinol. Metab*., **1**, 340–344.
7. Deretic,D., Schmerl,S., Hargrave,P.A., Arendt,A. and McDowell,J.H. (1998) Regulation of sorting and post-Golgi trafficking of rhodopsin by its C-terminal sequence QVS(A)PA. *Proc. Natl Acad. Sci. USA*., **18**, 10620–10625.
8. Davey,N.E., Trave,G. and Gibson,T.J. (2011) How viruses hijack cell regulation. *Trends Biochem. Sci*., **3**, 159–169.
9. Neduva,V., Linding,R., Su-Angrand,I., Stark,A., de Masi,F., Gibson,T.J., Lewis,J., Serrano,L. and Russell,R.B. (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol*., **12**, e405.
10. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., Cuche,B.A., de Castro,E., Lachaize,C., Langendijk-Genevaux,P.S. and Sigrist,C.J. (2008) The 20 years of PROSITE. *Nucleic Acids Res*., **36**, D245–D249.
11. Rajasekaran,S., Balla,S., Gradie,P., Gryk,M.R., Kadaveru,K., Kundeti,V., Maciejewski,M.W., Mi,T., Rubino,N., Vyas,J. *et al*. (2009) Minimotif miner 2nd release: a database and web system for motif search. *Nucleic Acids Res*., **37**, D185–D190.
12. Obenauer,J.C., Cantley,L.C. and Yaffe,M.B. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*., **13**, 3635–3641.
13. Puntervoll,P., Linding,R., Gemund,C., Chabanis-Davidson,S., Mattingsdal,M., Cameron,S., Martin,D.M., Ausiello,G., Brannetti,B., Costantini,A. *et al*. (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res*., **13**, 3625–3630.
14. Gould,C.M., Diella,F., Via,A., Puntervoll,P., Gemund,C., Chabanis-Davidson,S., Michael,S., Sayadi,A., Bryne,J.C. *et al*. (2010) ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res*., **38**, D167–D180.
15. Davey,N.E., Edwards,R.J. and Shields,D.C. (2010) Computational identification and analysis of protein short linear motifs. *Front. Biosci*., **15**, 801–825.
16. Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. *et al*. (2004) The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat. Biotechnol*., **2**, 177–183.
17. Cote,R.G., Jones,P., Apweiler,R. and Hermjakob,H. (2006) The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*., **7**.
18. Via,A., Gould,C.M., Gemund,C., Gibson,T.J. and Helmer-Citterich,M. (2009) A structure filter for the Eukaryotic Linear Motif Resource. *BMC Bioinformatics*., **10**.
19. Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res*., **37**, D229–D232.
20. Dosztanyi,Z., Csizmok,V., Tompa,P. and Simon,I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **16**, 3433–3434.
21. Dinkel,H. and Sticht,H. (2007) A computational strategy for the prediction of functional linear peptide motifs in proteins. *Bioinformatics*, **24**, 3297–3303.
22. Chica,C., Labarga,A., Gould,C.M., Lopez,R. and Gibson,T.J. (2008) A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. *BMC Bioinformatics*, **9**.
23. Balagopalan,L., Coussens,N.P., Sherman,E., Samelson,L.E. and Sommers,C.L. (2010) The LAT story: a tale of cooperativity, coordination, and choreography. *Cold Spring Harb. Perspect. Biol*., **8**, a005512.
24. Pawson,T. and Scott,J.D. (2005) Protein phosphorylation in signaling–50 years and counting. *Trends Biochem. Sci*., **6**, 286–290.
25. Dinkel,H., Chica,C., Via,A., Gould,C.M., Jensen,L.J., Gibson,T.J. and Diella,F. (2011) Phospho.ELM: a database of phosphorylation sites–update 2011. *Nucleic Acids Res*., **39**, D261–D267.
26. Cruz,J.L., Sola,I., Becares,M., Alberca,B., Plana,J., Enjuanes,L. and Zuniga,S. (2011) Coronavirus gene 7 counteracts host defenses and modulates virus virulence. *PLoS Pathog*., **6**, e1002090.
27. Kadaveru,K., Vyas,J. and Schiller,M.R. (2008) Viral infection and human disease–insights from minimotifs. *Front. Biosci*., **13**, 6455–6471.
28. Weil,D., El-Amraoui,A., Masmoudi,S., Mustapha,M., Kikkawa,Y., Laine,S., Delmaghani,S., Adato,A., Nadifi,S., Zina,Z.B. *et al*. (2003) Usher syndrome type I G (USH1G) is caused by mutations in the gene encoding SANS, a protein that associates with the USH1C protein, harmonin. *Hum. Mol. Genet*., **5**, 463–471.
29. Tapia,V.E., Nicolaescu,E., McDonald,C.B., Musi,V., Oka,T., Inayoshi,Y., Satteson,A.C., Mazack,V., Humbert,J., Gaffney,C.J. *et al*. (2010) Y65C missense mutation in the WW domain of the Golabi-Ito-Hall syndrome protein PQBP1 affects its binding activity and deregulates pre-mRNA splicing. *J. Biol. Chem*., **25**, 19391–19401.
30. Pandit,B., Sarkozy,A., Pennacchio,L.A., Carta,C., Oishi,K., Martinelli,S., Pogna,E.A., Schackwitz,W., Ustaszewska,A., Landstrom,A. *et al*. (2007) Gain-of-function RAF1 mutations cause Noonan and LEOPARD syndromes with hypertrophic cardiomyopathy. *Nat. Genet*., **8**, 1007–1012.
31. von Nandelstadh,P., Ismail,M., Gardin,C., Suila,H., Zara,I., Belgrano,A., Valle,G., Carpen,O. and Faulkner,G. (2009) A class III PDZ binding motif in the myotilin and FATZ families binds enigma family proteins: a common link for Z-disc myopathies. *Mol. Cell Biol*., **3**, 822–834.
32. Selcen,D. and Engel,A.G. (2004) Mutations in myotilin cause myofibrillar myopathy. *Neurology*, **8**, 1363–1371.
33. Gfeller,D., Butty,F., Wierzbicka,M., Verschueren,E., Vanhee,P., Huang,H., Ernst,A., Dar,N., Stagljar,I., Serrano,L. *et al*. (2011) The multiple-specificity landscape of modular peptide recognition domains. *Mol. Syst. Biol*., **7**.
34. Bauer,D.C., Willadsen,K., Buske,F.A., Le Cao,K.A., Bailey,T.L., Dellaire,G. and Boden,M. (2011) Sorting the nuclear proteome. *Bioinformatics*, **13**, i7–i14.
35. Walsh,I., Martin,A.J., Di Domenico,T., Vullo,A., Pollastri,G. and Tosatto,S.C. (2011) CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. *Nucleic Acids Res*., **39**, W190–W196.
36. Lieber,D.S., Elemento,O. and Tavazoie,S. (2010) Large-scale discovery and characterization of protein regulatory motifs in eukaryotes. *PLoS One*, **12**, e14444.
37. Pless,O., Kowenz-Leutz,E., Dittmar,G. and Leutz,A. (2011) A differential proteome screening system for post-translational modification-dependent transcription factor interactions. *Nat. Protoc*., **3**, 359–364.

38. Goel,R., Muthusamy,B., Pandey,A. and Prasad,T.S. (2011) Human protein reference database and human proteinpedia as discovery resources for molecular biotechnology. *Mol. Biotechnol.*, **1**, 87–95.
39. Edwards,R.J., Davey,N.E. and Shields,D.C. (2007) SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS One*, **10**, e967.
40. Edwards,R.J., Davey,N.E. and Shields,D.C. (2008) CompariMotif: quick and easy comparisons of sequence motifs. *Bioinformatics*, **10**, 1307–1309.
41. Perkins,J.R., Diboun,I., Dessailly,B.H., Lees,J.G. and Orengo,C. (2010) Transient protein-protein interactions: structural, functional, and network properties. *Structure*, **10**, 1233–1243.
42. Edwards,R.J., Davey,N.E., Brien,K.O. and Shields,D.C. (2011) Interactome-wide prediction of short, disordered protein interaction motifs in humans. *Mol. Biosyst.*, August 30 (doi:10.1039/c1mb05212h; epub ahead of print).